On statistical testing and mean parameter estimation for zero–modification in count data regression

Paul Wilson[1], Jochen Einbeck[2]

[1] School of Mathematics and Computer Science, University of Wolverhampton, UK
[2] Department of Mathematical Sciences, Durham University, UK

E-mail for correspondence: pauljwilson@wlv.ac.uk

Abstract: For the problem of testing for zero–modification in Poisson regression, a simple and intuitive test can be constructed by computing directly confidence intervals for the number of 0's under the Poisson assumption. This requires the ability of estimating the mean function accurately even if the data are in fact zero–inflated or deflated. A novel hybrid estimator is introduced for this purpose, which is of interest beyond the scope of the motivating test problem.

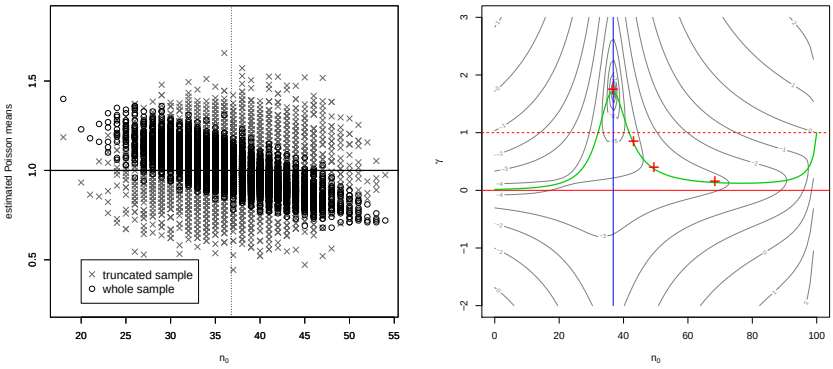
Keywords: Zero–modification; zero–truncated model; hypothesis testing

1 Introduction

Commonly used tests for zero–inflation/modification are likelihood ratio, score and Wald tests. Whilst these tests are all viable, they are not readily understood by non–statisticians, they do not distinguish between zero–inflation and zero–deflation (at least, not without adjustments), and they rely upon asymptotic results. Wilson and Einbeck (2015) proposed a new family of tests to test zero–modification in count data regression. Consider data $(y_i, \boldsymbol{x}_i^T), i = 1, \ldots, n$, where y_i are discrete counts and $\boldsymbol{x}_i \in \mathbb{R}^d$ a predictor vector. Let $p_i = P(y_i = 0)$. In the special case of (possibly zero–modified) Poisson regression, this test can be summarized as follows. For given significance level α: (i) fit the Poisson regression model, yielding Poisson means $\hat{\mu}_i = \hat{E}[y_i|\boldsymbol{x}_i]$; (ii) for each each y_i estimate $\hat{p}_i = \exp(-\hat{\mu}_i)$; (iii) use a Poisson–Binomial distribution with parameters $(n, \hat{p}_1, \ldots, \hat{p}_n)$ to determine a $1–\alpha$ confidence interval for the number of 0's.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016.

FIGURE 1. Left: Estimation from the zero-truncated and whole sample; right: Function $\gamma^*_{100}(n_0, 1)$ (thick curve) with $MSE(T|n_0)$ contours. In both plots, $\mu = 1$ and $n = 100$.



The challenging part in this procedure is the estimation of the Poisson means $\mu_i = E[y_i|\boldsymbol{x}_i]$ in the absence of the knowledge whether the Poisson assumption is correct. This problem has attracted attention earlier; Dietz and Böhning (2001) observed that ML estimation of the zero–modified Poisson model can be obtained by ML estimation of the zero–truncated Poisson (ZTP) model. For additional insight, consider Figure 1 (left), which shows the estimates of the Poisson means obtained when $n = 100$ observations are sampled from a Pois(1) distribution. The black circles indicate whole sample mean (Poisson) estimates $\hat{\mu}_P$, and the grey crosses the means $\hat{\mu}_T$ obtained from the positive observations. The horizontal axis gives the number of zeros, n_0, with the expected number of zeros under the Poisson model, $100e^{-1} \approx 37$, highlighted by a dotted line. It is clear that the Poisson estimator has smaller variance but is possibly biased if the observed number of zeros is far from their expected number. On the other hand, the ZTP–derived mean estimator does not demonstrate a noticeable bias, at the expense of a large variance.

2 A hybrid mean estimator

The illustrated bias–variance trade–off motivates the definition of the hybrid estimator

$$T = \gamma\hat{\mu}_P + (1 - \gamma)\hat{\mu}_T \tag{1}$$

which is a weighted sum of the usual Poisson mean estimator $\hat{\mu}_P$ and an estimator of the zero–truncated mean, $\hat{\mu}_T$. The latter is based on the mean of the zero–truncated data only, to which we refer from now on as ζ. Note that the mean μ of a Poisson distribution and the mean ζ of the ZTP

distribution are related by $\zeta = \frac{\mu e^{\mu}}{e^{\mu}-1} \equiv h(\mu)$. The MLE of μ under the ZTP assumption is then given by the inverse mapping $\hat{\mu}_T = h^{-1}(\hat{\zeta})$. Of course, all terms used in this section can be equipped with the index i to account for the case of covariates as laid out in Section 1.

3 Selection of the hybrid parameter

For the choice of γ, we have initially carried out a detailed theoretical study. To give some idea of this, we provide here the result that, in the covariate–free case, and only assuming a ZTP distribution for the non–zero part, the $MSE(T|n_0)$ is minimized at

$$\gamma_n^*(n_0, \mu) = \frac{\frac{n}{n-n_0} - h'(\mu)}{\frac{1}{n}\frac{\mu(n-n_0e^{\mu})^2 h'(\mu)^2}{e^{\mu}(e^{\mu}-1-\mu)} + h'(\mu)^2\left(1-\frac{n_0}{n}\right) + 2h'(\mu) + \frac{n}{n-n_0}} \qquad (2)$$

Figure 1 (right) shows the curve γ^* for fixed $n = 100$ and $\mu = 1$. It is, firstly, interesting to note that in a small range close to the expected value (≈ 37) under the Poisson model, the optimal γ is in fact > 1. However, for the majority of values of n_0 the curve is between 0 and 1, and falls very quickly below 1 when deviating from the expected value. While this kind of result could motivate an iterative procedure, in which T and γ are updated in turns via (1) and (2), we found this approach practically less useful since the increased variance incurred by the iterative estimation of γ contravenes the purpose of the hybrid estimator. We therefore considered two considerably simpler schemes:

(i) a single fixed rule–of–thumb value; where we have chosen $\gamma = 2/3$.

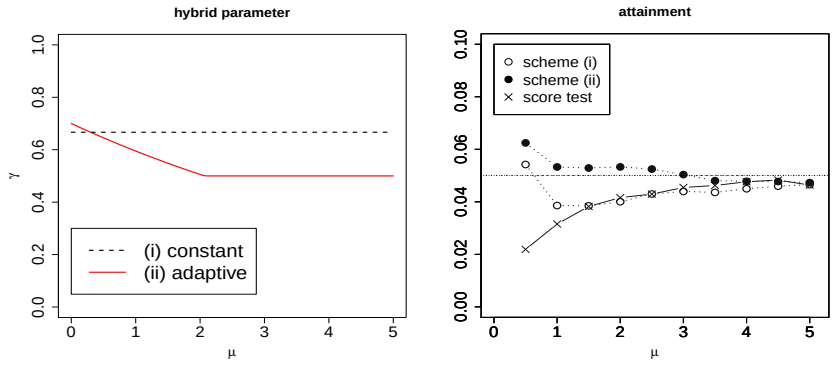
(ii) a parametric expression $\gamma = f(\hat{\mu}_P) = \begin{cases} 0.7\left(0.85^{\hat{\mu}_P}\right) & \hat{\mu}_P < \frac{\log(5/7)}{\log(17/20)} \\ \frac{1}{2} & \textit{otherwise} \end{cases}$

The rationale of (ii) is to improve the attainment rate of the test by dereasing the weighting of the Poisson mean in the mixture for larger values of this estimator. The threshold $\frac{\log(5/7)}{\log(17/20)} \approx 2.07$ is chosen so that f is continuous. Figure 2 (left) compares settings (i) and (ii) graphically. Consider in this context the four crosses, from left to right in Figure 1 (right), which correspond to the optimal γ under zero–inflation parameter 0, 0.1, 0.2 and 0.5, respectively. We see that in the middle two cases (moderate zero–inflation) one has $\gamma^* \in [0.4, 0.8]$, so that we consider our suggested choices to be in harmony with our theoretical considerations.

4 Simulation

For the two–sided zero–modification test, Figure 2 (right) demonstrates, for a covariate–free simulation from Poisson data of varying μ, that (i) and

FIGURE 2. Left: choices (i) and (ii) for the selection of γ; right: attainment rate under mixture estimator (Two sided test of zero-modification)



(ii) both work well in terms of the nominal level attainment, with slight advantages for (ii). Focusing now on (ii), Figure 3 gives an impression of power as compared to the score test, as a function of sample size n. One sees that the powers are strong and very competitive to the score test, especially for smaller sample sizes. Note that here, and throughout this paper, the p-values reported for the proposed test are the mid p-values $\frac{1}{2}P_0[T \geq t+1] + \frac{1}{2}P_0[T \geq t]$ of Franck (1986).

FIGURE 3. Power under mixture estimator (Covariate–free Model, Two sided test of zero–modification)

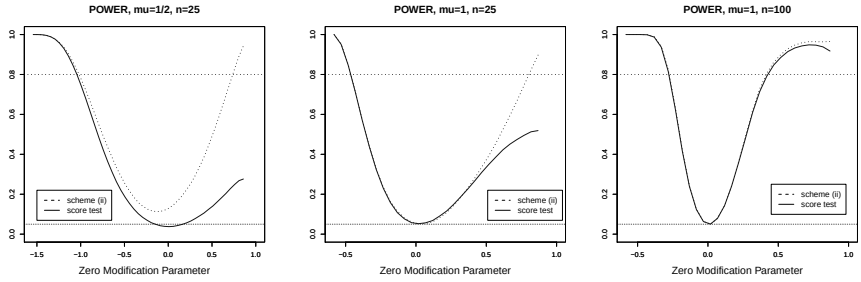
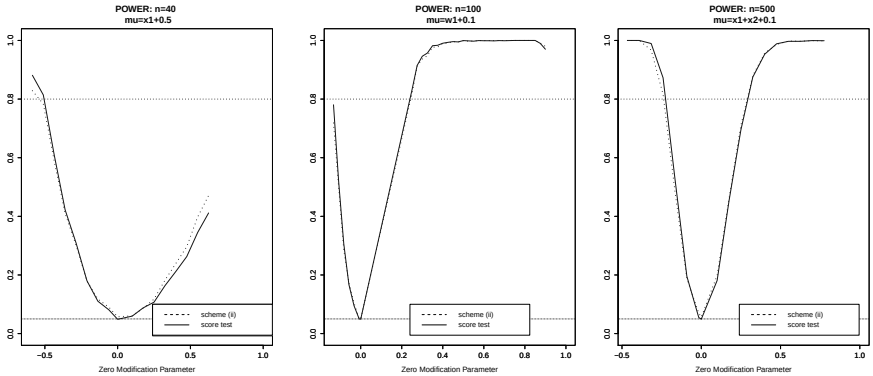


Figure 4 shows that the power and nominal attainment level of the proposed test also compares strongly to that of the score test in the presence of covariates. Here x_1 and x_2 are uniformly distributed on the interval $(0, 0.5)$, and w_1 is uniformly distributed on the interval $(1, 2)$. The adaptive mixing parameter is used, but the results remain similar for the constant estimator.

FIGURE 4. Power under mixture estimator (Covariate Model, Two sided test of zero–modification)



5 Examples

5.1 Biodosimetry Data

We consider four biodosimetry datasets consisting of chromosome aberration counts occuring after whole body exposure to ionising radiation. These datasets have previously been studied by Oliveira et al. (2016), detailed descriptions of the datasets are available in this paper.
Table 1 summarises the results obtained when the proposed test and a score test are used to test for zero-inflation relative to a quadratic Poisson model with log-link. We see that both tests fail to reject the Poisson model for the A3 data, but do not do so for the other datasets considered. For all the instances where the Poisson model was rejected we see that the observed number of zeros is greater than the upper limit of the 95% confidence interval, indicating that the data is zero-inflated.

TABLE 1. Analyses of Biodosimetry Data

	Proposed Test			Score Test	
Data	Obs. Zeros	$95\%CI$	p-value	Statistic	p-value
A1	$14,430$	$(14204, 14329)$	$< 10^{-9}$	16.85	4.03×10^{-5}
A3	$2,747$	$(2719, 2823)$	0.368	1.01	0.317
B1	$7,280$	$(6707, 6829)$	$< 10^{-9}$	87.16	$< 10^{-9}$
C1	$6,786$	$(5031, 5164)$	$< 10^{-9}$	$1,996.10$	$< 10^{-9}$

5.2 Unwanted Pursuit Behaviour Data

Loeys et al. (2012) analysed data which concerns "separation trajectories". Participants in a survey were assigned a score that theoretically ranges from

0 to 112, the maximum observed score was 34. This score is a measure of the participants experience of behaviour by their partner that contributed towards the breakup of a relationship. Two covariates were included in the model: a binary variable "education level" (0 = lower than bachelors degree, 1 = at least bachelors degree), and a continuous measurement for the level of anxious attachment in the former partner relationship. There are $n = 387$ data of which 246 are zeros. The proposed test shows that a 95% confidence interval for the number of observed zeros under the Poisson model is $(45, 72)$, and hence we may reject the Poisson model. Analysis of the data by a score test returns a statistic of 591.8, also indicating rejection of the Poisson model. Both tests return p-values $< 10^{-9}$.

6 Conclusion

The proposed test for zero-modification has power and attainment rates that compare very strongly to the score test. In addition to this it distinguishes between zero-inflation and zero-deflation and is a highly intuitive test that, unlike existing tests, is readily explainable to non-statistical specialists. The technique may be extended to compare any two count regression models, and may be used as the basis of a diagnostic plot for assessing model fit. See Einbeck and Wilson (2016).

References

Dietz, E. and Böhning, D. (2000). On estimation of the Poisson parameter in zero–modified Poisson models. *Computational Statistics & Data Analysis* **34**, pages 441 – 459.

Einbeck, J. and Wilson, P. (2016). A Diagnostic Plot for Assessing Model Fit. Proc's of the 31st IWSM, Rennes, France, *to appear.*

Franck, W. (1986). P-values for Discrete Test Statistics. *Biometrical Journal* **4**, pages 403 – 406.

Loeys,T., Moerkerke, B., De Smet, O., and Buysse, A. (2012) Expert Tutorial: The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology* **65**, pages 163 – 180.

Oliveira, M., Einbeck, J., Higueras, M., Ainsbury, E., Puig, P. and Rothkamm, K (2016). Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. *Biometrical Journal* **58**, 259 - 279.

Wilson, P. and Einbeck, J. (2015). A simple and intuitive test for number–inflation or number–deflation. In: Wagner, H. and Friedl, H. (Eds). Proc's of the 30th IWSM, Linz, Austria, Vol 2, pages 299 – 302.